WOODHOUSE EXHIBIT 2

EXHIBIT C

Draft | Work in Progress | Review | **Published**

AI Data Catalog 2024H1 Roadmap

Part of 2024H1 planning for [HYPERLINK "https://docs.google.com/document/d/[redacted]edit" \l "heading=h.kf98c991lcvi" \h] and [HYPERLINK "https://docs.google.com/document/d/[redacted]edit" \l "heading=h.87so7gpd6f5m" \h]
Docs: [HYPERLINK "https://docs.google.com/document/d/[redacted]edit" \h], [HYPERLINK "https://docs.google.com/document/d/[redacted]edit" \l "heading=h.t6rlu5nbxv2y" \h]
WP groups: [HYPERLINK "https://fb.workplace.com/groups/[redacted]" \h], [HYPERLINK "https://fb.workplace.com/groups/[redacted]" \h]

People: [HYPERLINK "mailto:[redacted]@meta.com" \h] (Planning STO),[HYPERLINK "mailto[redacted]@meta.com" \h], [HYPERLINK "mailto:[redacted]@meta.com" \h], [HYPERLINK "mailto:[redacted]@meta.com" \h]

Vision - the purpose of the AI Data Catalog

The AI Data Catalog provides a unified experience for all data used in AI workflows: sourcing datasets, training data, evals and flywheels, covering 1st and 3rd party data, as well as synthetic data. It serves as the single source of truth for data discovery, sourcing, management, evaluation and governance.

It enables responsible and compliant use of data, through Privacy Review integration and by keeping track of and enforcing lineage, data mitigation and curation metadata, privacy and legal policies.

Integrating data insights, visualizations, eval results, model shortcomings and sourcing in a unified experience speeds up data decisions, augmentation and curation, and therefore speeds up model performance improvements.

The AI Data Catalog integrates with systems throughout the entire AI Workflow execution, from authoring (Dataswarm Operators, Bento) to data transformation (DPS) to training (Genie) to evaluation (Halo), ensuring each step has an out-of-the-box or programmatic way of registering metadata and enforcing policies.

How Success Looks Like

1. It's easy to understand **what data was used** to train which model and which mitigations have been applied to the datasets so **PXFN approval** is a lightweight process
2. The data catalog is the default **place of choice** for researchers looking for datasets for new ML tasks.
3. Dataset owners can spend more time on improving datasets or creating new ones instead of handling routine dataset **lifecycle tasks** which are now **automated** and centrally managed.

OKRs

OKRs	**H1'24 target (p50)**
Objective1: Improve datasets compliance	
AIDC1.1 % of datasets registered in AIDC	100% of GenAI (includi e
AIDC1.2 % of datasets with lineage to models	100% of GenAI datase
AIDC1.3 % of existing Hive training datasets with dataset to features lineage	X% (for datasets that contain features)
Objective2: Reduce the overhead of datasets privacy reviews	
AIDC2.1 % of AIDC datasets with PXFN enabling metadata	[HYPERLINK "mailto:[redacted]@meta.com" \h]
AIDC2.2 % of existing mitigations (priv reqs) and approvals available in AIDC	TBD [HYPERLINK "mailto:[redacted]@meta.com" \h]
AIDC2.3 # of LaMas with pre-approved datasets	[HYPERLINK "mailto:[redacted]@meta.com" \h]
Objective3 Help dataset owners manage their datasets	
AIDC3.1 MAU of GenAI AIDC users	50
AIDC3.2 Monthly retention rate of GenAI dataset owners	45%
Objective4 Help researchers improve models with better data	
AIDC4.1 MAU of GenAI AIDC users	50 (dup of 3.1)
AIDC4.2 Monthly retention rate of GenAI researchers	45%
AIDC4.3 # data sets registered through AIDC for HALO	Pending alignment w Halo [HYPERLINK "mailto:[redacted]@meta.com"
AIDC4.4 % of Genie data configs is generated by AIDC	Pending alignment w Genie [HYPERLINK "mailto[redacted]@meta.com
AIDC4.5 # of Dataswarm pipelines using AIDC datasets annotations	Pending Alignment with Xuchao [HYPERLINK "mailto:[redacted]@meta
AIDC4.6 # of LLM researchers consuming datasets Insights in AIDC	20 (10 P90)
Objective5 Supports the required reliability requirements to support inline training jobs	
AIDC5.1 AIDC SLA is 99.9%	99.9%
AIDC5.2 AIDC and Damit Unified UI visualization strategy	Design Doc

Commented [1]: Pre-training: Llama datasets
Fine tuning + flywheel: Genie
Eval: TBD

Commented [2]: @[redacted]@meta.com

Projects

Priorities P0: Must-have • P1: Want-to-have • P2: Nice-to-have.
T-Shirt Size: S: < 2 weeks • M: 2~4 weeks • L: 4~8 weeks • XL : > 8 weeks

Priority	OKR	Project	Description	Owning Team	Partner Teams	T-S
P0	3.1, 4.1	[HYPERLINK "https://docs.google.com/document/d/[redacted]edit" \h]	Separate datasource access control from metadata access control and allow dataset owners to manage the access	Damit	Damit DI Privacy	
P0	1.1	Datasets [HYPERLINK "https://docs.google.com/document/d/[redacted]/edit" \l "heading=h.t6rlu5nbxv2y" \h]	Onboard remaining teams and datasets: Speech, Audio/Music, Video, etc...	AIDC		
P0	2.1, 2.2	PXFN Support - [HYPERLINK "https://docs.google.com/document/d/[redacted]edit" \l "heading=h.p2t5dehvrwpo" \h]	Simplify facts gathering process for PR by allowing datasets owners to store facts in a structured way. Integrate with PR2.0	AIDC	PR2.0 DI Privacy FAIR Capella	
P0	1.1	[HYPERLINK "https://docs.google.com/document/d/[redacted]edit" \h]	Support composite datasets (LLM next) management, privacy and visualization	AIDC	LLM Pre-Training	
P0	4.4, 5.1	[HYPERLINK "https://docs.google.com/document/d/[redacted]edit" \h]	Integrate with the Genie training platform as the dataset management solution (configuration, discovery, resolve latest version, etc.)	AIDC	Genie	
P0	4.3	[HYPERLINK "https://docs.google.com/document/d/[redacted]	Halo datasets creation discovery and lineage.	AIDC	Halo	

		[redacted]edit" \l "heading=h.1rhahvrssxn5" \h]			
P0	1.2	[HYPERLINK "https://docs.google.com/document/u/0/d/1-[redacted]/edit" \h]	Support lineage display and automatic fetching of non ds partitioned based datasets + define how to propagate metadata	AIDC	CU
P0	1.1	Automated lineage for datasets in AirStore/RSC/EAG	LLM3 models are trained on RSC & EAG would like to be able to automatically sur lineage for assets outside of prod		
P0	4.1, 4.2	[HYPERLINK "https://docs.google.com/document/u/0/d/[redacted]edit" \h]	Enable configuration for visualization required pre-processing in AIDC (Data schematization, algorithm selection)	AIDC	Visualization, Evaluation and data insights
P0	2.2	Fetching [HYPERLINK "http://mitigations" \h] and [HYPERLINK "https://docs.google.com/document/d/[redacted]/edit" \h]	Enabling presentations of as many existing mitigations and approval related to each dataset/datasource as possible, including annotations from GDA	AIDC	DI Privacy PR2.0 FAIR
P0	1.1	Python API	Enabling users to register datasets using python (i.e in Bento notebooks, scripts)	AIDC	LLM Flywheel
P1	1.1, 1.2	[HYPERLINK "https://docs.google.com/document/d/[redacted]/edit" \h]	Unified configurable lineage view to display both datasets to datasets, datasets to model snapshots and model checkpoints to reuse across AI and DI	AIDC	AIM
P1	4.5	[HYPERLINK "https://docs.google.com/document/	Register datasets directly into AIDC from tool like Dataswarm (P0), Daiquery and Bento	AIDC	DxI Dataswarm

Commented [3]: @[redacted]@meta.com Do not have access to the one pager, commenting here. We should have a mode where automation is turned off for a dataset, to enable programmatic control of dataset versioning / lineage. This will be needed for the truly custom management. _Assigned to [redacted]@meta.com_

		d/[redacted]/edit" \h]				
P1	3.1, 3.2	[HYPERLINK "https://docs.google.com/document/d/[redacted]edit" \l "heading=h.gkoow11g2neo" \h]	Life cycle management including deletion syncing, dataset regeneration, retention alerting, AIDC Bot etc.	AIDC	DI Privacy Metastore Gaid	
P1	4.1	[HYPERLINK "https://docs.google.com/document/d/[redacted]edit" \h]		AIDC	CU	
P1	*	Innovation bucket	Based on existing [HYPERLINK "https://docs.google.com/spreadsheets/d/[redacted]edit" \l "gid=1587878533" \h] and emerging pain points and opportunities	AIDC		
P2	4.1, 4.2	Search and Discovery	Extend search and discovery to additional fields and federated metadata, add sorting and datasets count.	AIDC		
P2	1.1	Support auto-registered dataset	Currently we register datasets automatically for lineage purposes. These datasets have no metadata, we'd like to define propagation policies and let users control auto registered datasets	AIDC		
P2	5.1	AIDC types search scraper & bulk indexers	Improve search reliability by introducing daily scraping of AIDC assets	AIDC		
P2	5.1	Damit search unification	Unify AIDC & Damit search stack	AIDC		

Dependencies We Have on Others

Partner	Level	Description	AIDC KRs	AIDC POC
DI AIM	High	[HYPERLINK "https://docs.google.com/document/d/[redacted]/edit" \h]: KR 3.1: Extend AIM support to the EAG to register all EAG trained models and training pipeline to AIM [HYPERLINK "https://docs.google.com/document/d/[redacted]edit" \h]		[HYPERLINK "mailto:[redacted]@meta.con
ULP/HALO	High	[HYPERLINK "https://docs.google.com/document/d/[redacted]edit" \l "heading=h.zajqawcvdu2y" \h]: Integrate with AIDC: X data sets registered through AIDC for HALO [HYPERLINK "https://docs.google.com/document/d/[redacted]/edit" \l "heading=h.1rhahvrssxn5" \h]		[HYPERLINK "mailto[redacted]@meta.com"
DAMIT	High	ACL v2. Critical path reliability		[HYPERLINK "mailto[redacted]@meta.com"
DI Eval & Insights	High	Consumption of insights in AIDC[HYPERLINK "https://docs.google.com/document/d/[redacted]edit" \l "heading=h.biesry7lmmn2" \h]	4.1	[HYPERLINK "mailto[redacted]@meta.cor
Central Privacy	Med	PR2.0		[HYPERLINK "mailto[redacted]@meta.com" HYPERLINK "mailto[redacted]@meta.com" \h
Privacy GDA	Med			[HYPERLINK "mailto[redacted]@meta.com" HYPERLINK "mailto[redacted]@meta.com" \h
MLHub	Low	[HYPERLINK "https://docs.google.com/document/d/[redacted]/edit" \l "heading=h.qtk6ug8ecdfo" \h]		[HYPERLINK "mailto[redacted]@meta.com"
DI Bento/Daiquery	Low		1.1	[HYPERLINK "mailto:[redacted]@meta.c
DI Datawarm	Low		4.5	[HYPERLINK "mailto[redacted]@meta.con

Dependencies Others Have on Us

Partner	KR

Commented [4]: @[redacted]@meta.com, should we also add DAI for GenAI Privacy here for each reference by considering it has multiple AIDC dependencies: https://docs.google.com/document/d/[redacted]edit?
cc: @[redacted]@meta.com
Assigned to [redacted]@meta.com

GenAI Platform [HYPERLINK "https://docs.google.com/document/d/[redacted]edit" \h]	**3.1:** 100% of production models with full lineage tracked in system	1.2
	3.3: 50%+ of new production models using catalog	1.1
GenAI Platform Privacy and Safety [HYPERLINK "https://docs.google.com/document/d/[redacted]/edit" \l "heading=h.yc79vsiqsz30" \h], [HYPERLINK "https://docs.google.com/spreadsheets/d/[redacted]edit" \l "gid=675807207" \h]	**Purpose limitation 2.1:** GenAI models have known end-to-end lineage coverage for fine-tuning and pre-training, with [25%] done via automated coverage	1.1, 1.2
	Privacy 3.3: Ad hoc dataset reviews take less than one week and approved datasets are registered and labeled in AIDC	2.1, 2.2, 2.3
	Privacy 4.1: Make PXFN reviews for generative AI products on the platform more efficient cutting review time by 50%	2.1, 2.2, 2.3
	IP 3.3: Ad hoc dataset reviews take less than one week and approved datasets are available in AIDC	2.1, 2.2, 2.3
ULP/Halo [HYPERLINK "https://docs.google.com/document/d/[redacted]edit" \l "heading=h.zajqawcvdu2y" \h]	**Platform Growth and Maturity:** X data sets registered through AIDC for HALO	4.3
GenAI Platform Genie Model Hub [HYPERLINK "https://docs.google.com/document/d/[redacted]edit" \h], [HYPERLINK "https://docs.google.com/document/d/[redacted]/edit" \h]	**KR2.1.1:** % of Genie configs with datasets registered in AIDC **KR2.1.2:** % of Genie data configs is generated by AIDC	4.4
FAIR [HYPERLINK "https://docs.google.com/document/d/[redacted]edit" \h], [HYPERLINK "https://docs.google.com/spreadsheets/d/[redacted]edit" \l "gid=0" \h]	**4.1** Import 100% of FAIR datasets from the canonical spreadsheet and existing data sources to the dedicated Data Catalog (Q1)	1.1
	4.2 Achieve near-zero dedicated review for top tier CV and S&A datasets, through integration of Data Catalog with Privacy Review pilot stored decisions	2.1, 2.2, 2.3
	4.3 Reduce the time of fact gathering during PXFN reviews for CV and S&A datasets by x%	2.1, 2.2

AIM [HYPERLINK "https://docs.google.com/document/d[redacted]edit" \h]	Depend on AIDC API to auto curate AIDC dataset when AIM capture the physical dataset, and build the lineage from model to AIDC dataset	1.1, 1.2
DPS [HYPERLINK "https://fb.workplace.com/notes/[redacted]?notif_id=1704311018108105&notif_t=work_garden_note_mention&ref=notif" \h]	Ingest into DPS via AIDC, with Amnesia 2.0 support. Make sure data is registered with AIDC. Generate profiling reports through AIDC, both automatic and manual options are available.	4.4

Open Decisions

	Status	Next steps/Decision
DAMIT and what is our approach to new customers onboarding such as FAIR?	Resolved	DAMIT and AIDC are on a convergence path, we AIDC and vice versa. Both teams desire to work t
GenAI roadmap?	Resolved	Rationalize the diff GenAI Platform roadmaps and [redacted]edit"
vners and MLEs?	In progress	

Risks

isk	Level	Description	Mitigation
rivacy in Flux	High	Privacy requirements are in flux and keep changing. e.g . The fact gathering and meta data store for mitigations and evidence	
lany dependencies on external eams	High	Need to bring together multiple teams, agree on priorities, tech strategy and sync on roadmap items	Create transparency and document each dependency Communicate frequently and publicly Establish dedicated WP chats/groups
ataset Metadata Completeness	Med	Define the process changes that are required to register datasets and the associated metadata and mitigations	
romoting the AI Dataset concept	Med	How to ensure the definition and adoption of the AI dataset primitive across different systems and configurations so we're on top of our compliance requirements from day 1	